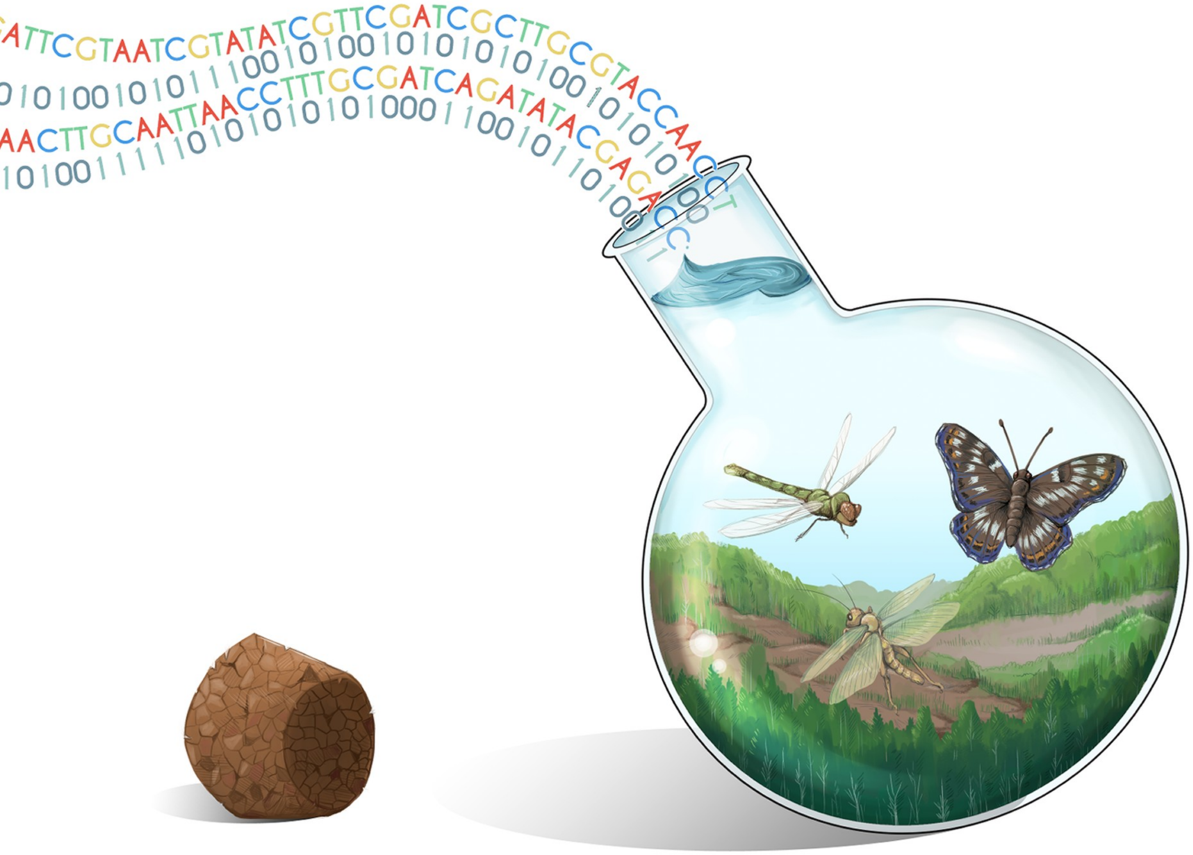
Stockholm
University

Debugging metabarcoding for insect biodiversity studies

Daniel Marquina Hernández



Doctoral Thesis in Systematic Zoology at Stockholm University, Sweden 2020

Debugging metabarcoding for insect biodiversity studies

Daniel Marquina Hernández

Academic dissertation for the Degree of Doctor of Philosophy in Systematic Zoology at Stockholm University to be publicly defended on Friday 15 May 2020 at 09.30 in Vivi Täckholmsalen (Q-salen), NPQ-huset, Svante Arrhenius väg 20.

Abstract

Insects are one of the most abundant and diverse animal groups, and they include many valuable ecological indicator species, but taxonomic discovery projects and biodiversity surveys targeting this group are often challenging. While mass trapping devices allow the collection of insects in great numbers, the task of identifying the species present is a painstaking and resource-demanding process. Metabarcoding, that is, high throughput sequencing of PCR-amplified species-specific genetic markers in environmental samples, promises to solve this problem. However, metabarcoding is still in its infancy. In this thesis, I optimized metabarcoding methods for inventorying and accelerating species discovery of terrestrial insects. In **paper I,** we designed new PCR primers for mitochondrial markers and evaluated them against existing ones using *in silico* methods. We showed that the best marker for metabarcoding of insects is 16S because of its broad taxonomic coverage and low amplification bias. However, there is significantly more reference data for COI, and its taxonomic coverage is reasonable when using sufficiently degenerate primers (mixes of primer sequences). In **paper II,** we applied 16S and COI metabarcoding to different types of samples of the same insect communities: Malaise trap samples (preservative ethanol or homogenized samples) and soil samples. The results show that the two-marker strategy increases biodiversity detection over single-marker analyses. They also show that 16S is better than COI for metabarcoding of eDNA samples because the less degenerate 16S primers do not amplify as many off-target organisms. Finally, the results show that analyses of tissue homogenate and preservative ethanol yield strikingly different results. Large and heavily sclerotized insects do not leak DNA into preservative ethanol like small and weakly sclerotized ones do, but their DNA tends to swamp the DNA of the latter in homogenized samples. In **paper III** we evaluated the performance of various non-destructive mild lysis treatments and DNA purification methods. We subjected mock community samples to incubation in either a milder or a more aggressive digestion buffer for a short or a long period of incubation. The DNA was then extracted using either a manual or an automated purification protocol. We found that the milder digestion buffer and the shorter incubation time preserved the morphology of the insect best while at the same time giving the most accurate DNA metabarcoding results; the purification protocol had little or no effect on metabarcoding results. Finally, in **paper IV,** we explored the received wisdom that high concentrations of ethanol, although optimal for preservation of the DNA, make insects fragile and difficult to work with from a morphological point of view. We preserved insects in different ethanol concentrations and subjected them to damaging processes, such as shaking or transporting. We verified that high concentrations of ethanol induce brittleness, although the effect is less pronounced in robust insects. Our results also indicate that shipping by mail is safe for samples preserved at intermediate concentrations (70 or 80 %). In summary, this thesis represents a significant step forward in the development of methods for preserving and analyzing samples of terrestrial insects for biodiversity surveys, monitoring programs, and taxonomic research projects.

Keywords: *metabarcoding, insects, non-destructive, DNA extraction, Malaise trap, preservative ethanol, environmental DNA.*

Stockholm 2020
http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-180293

ISBN 978-91-7911-130-4
ISBN 978-91-7911-131-1

Department of Zoology

Stockholm University, 106 91 Stockholm

DEBUGGING METABARCODING FOR INSECT BIODIVERSITY STUDIES

Daniel Marquina Hernández

Stockholm University

Debugging metabarcoding for insect biodiversity studies

Daniel Marquina Hernández

ISBN print 978-91-7911-130-4
ISBN PDF 978-91-7911-131-1

Cover artwork by Allison Hsiang

Printed in Sweden by Universitetsservice US-AB, Stockholm 2020

[...]
- I formed an idea and then discovered I was wrong.
- There are numerous diagrams.
- I was wrong in numerous ways. I produced a detailed tribute to my wrongness.
- That is science!

- Nathan W. Pyle -

The thesis is based on the following articles, which are referred to in the text by their Roman numerals:

I **Marquina, D.**, Andersson, A.F. & Ronquist, F. (2019). New mitochondrial primers for metabarcoding of insects, designed and evaluated using *in silico* methods. *Molecular Ecology Resources* 19(1), 90–104.

II **Marquina, D.**, Esparza-Salas, R., Roslin, T & Ronquist, F. (2019). Establishing arthropod community composition using metabarcoding: Surprising inconsistencies between soil samples and preservative ethanol and homogenate from Malaise trap catches. *Molecular Ecology Resources* 19(6), 1516–1530.

III **Marquina, D.**, Roslin, T., Łukasik, P. & Ronquist, F. Evaluation of non-destructive extraction protocols for metabarcoding of insects. *Manuscript*.

IV **Marquina, D.**, Ronquist, F. & Łukasik, P. (2019). The effect of ethanol concentration on the preservation of insects for biodiversity studies. *bioRxiv*, 2020.03.05.978288v1 (*Manuscript preprint*).

Candidate contributions to thesis articles*

	I	**II**	**III**	**IV**
Conceived the study	Significant	Substantial	Significant	Significant
Designed the study	Substantial	Substantial	Significant	Significant
Collected the data	Substantial	Substantial	Substantial	Substantial
Analysed the data	Substantial	Substantial	Substantial	Substantial
Manuscript preparation	Substantial	Substantial	Substantial	Substantial

*** Contribution Explanation**

Minor: contributed in some way, but contribution was limited.
Significant: provided a significant contribution to the work.
Substantial: took the lead role and performed the majority of the work.

The two published papers are open access and are herein reproduced under the course of the Creative Commons Attribution License (CC-BY).

During this period I have also co-authored the following articles:

Aguado, M. T., Noreña, C., Alcaraz, L, **Marquina, D.**, Brusa, F., Damborenea, C., Almon, B., Bleidorn, C. & Grande, C. (2017). Phylogeny of Polycladida (Platyhelminthes) based on mtDNA data. *Organisms Diversity & Evolution* 17(4), 767–778.

Matos-Maraví, P., Duarte-Ritter, C., Barnes, C. J., Nielsen, M., Olsson, U., Wahlberg, N., **Marquina, D.**, Sääksjärvi, I. & Antonelli, A. (2019). Biodiversity seen through the perspective of insects: 10 simple rules on methodological choices, common challenges, and experimental design for genomic studies. *PeerJ* 7, e6727.

CONTENTS

INTRODUCTION

In the current scenario of climate change and increasing human impact across all ecosystems at a global scale, the task of documenting biodiversity is of paramount importance. We need to fully understand diversity in order to effectively prevent major ecological damage caused by dwindling populations and the loss of species. Insects represent a large fraction of multicellular life but it is only recently that the full extent of the decline in their abundance is starting to be exposed (Hallmann et al., 2017; Lister & Garcia, 2018). The decline in insect populations can have serious impacts on ecosystem functions that we are just starting to discover (Lister & Garcia, 2018).

Whether insects are the focus of taxonomic research or ecological assessment, there is one critical bottle-neck: species identification. The traditional workflow of taxonomic research projects or biomonitoring campaigns consists of collection, sorting, species identification and analysis of the results. Collection is usually done passively with various types of traps (Malaise traps, pitfall traps, yellow pan traps, etc.) for those insects found in terrestrial environments, or actively by net sweeping or kick-sampling for those species or life stages that inhabit freshwater habitats. The sample is kept in a preservative fluid, *e.g.* ethanol, and sorted into smaller taxonomic fractions. The sorting requires time but it can be done by trained personnel who do not necessarily need to be experts in insect taxonomy. The great bottleneck of the process comes at the point of identification, as all individuals must be examined one by one. Also, this requires taxonomic expertise for each insect group examined, which is not always available or easily accessible.

New genetic identification methods have the potential to address this bottleneck. The identification of species based on a short fragment of a variable region of DNA is known as DNA barcoding (Hebert, Cywinska, Ball, & deWaard, 2003). DNA barcoding revolutionized how we identify species since it provided reliable identifications, in principle without dependence on taxonomic expertise once appropriate reference libraries had been constructed, while accelerating the process at the same time. The theory behind it is that, for most animals, individuals of the same species typically present a genetic distance (number of mismatches in a comparison of two DNA sequences) of at most 2 % in a 658 bp region at the 5' end of the cytochrome oxidase I (COI) gene, while the genetic distance for individuals of different species is larger than 2 % (Hebert, Ratnasingham, & de Waard, 2003). There is often a gap in the frequency distribution of genetic distances around this threshold, denominated the 'barcoding gap' (Meyer & Paulay, 2005). Thus, obtaining the sequence of this ‘barcoding region’ of COI – also known as ‘Folmer region’ as it spans the COI fragment that is amplified with the primers designed by Folmer and collaborators (Folmer, Black, Hoeh, Lutz, & Vrijenhoek, 1994) – would allow a researcher to identify a specimen by simply matching this barcode to a reference library. If the barcode falls on the intraspecific side of the barcoding gap for one of the reference sequences, it has been identified to species Apart from alleviating the lack of taxonomic expertise, DNA barcoding can solve cases of cryptic species, identify life stages or sexes without diagnostic characters, or simply greatly reduce the time of processing the samples (Hebert, Penton, Burns, Janzen, & Hallwachs, 2004; Janzen et al., 2009; Telfer et al., 2015). Does this mean that taxonomy is no longer needed? Quite the opposite. DNA barcoding benefits from taxonomy as much as taxonomy can benefit from DNA barcoding. The existence of a sequence database, well curated taxonomically speaking and globally accessible, is pivotal for the accuracy of DNA barcoding, and such a resource can only be created in collaboration with taxonomists covering all groups of organisms. The BOLD system represents the largest effort to date to create a global reference database for DNA barcoding (Hebert & Ratnasingham, 2007). Currently, BOLD contains more than 8 million barcodes belonging to approximately 670 thousand BINs, and almost 220 thousand animal species.

The development of high throughput sequencing (HTS) technologies in recent years has changed the field of DNA barcoding. It is now possible to simultaneously generate millions of sequences from a single DNA sample. If HTS sequencing is applied to a sample with a mix of PCR-amplified barcoding DNA from different species, this is called metabarcoding (Pompanon, Coissac, & Taberlet, 2011; Riaz et al., 2011; Taberlet, Coissac, Hajibabaei, & Riesenberg, 2012). Metabarcoding can be applied to any sample containing a mix of amplified DNA from different species. This includes samples that predominantly contain individuals of the target group of interest, such as a trap catch, or samples that contain only traces of DNA that is shed into the environment where the organisms live: soil, water, air or sediments. The former are often called bulk samples, while the latter are referred to as eDNA samples (for environmental DNA). The same sample can be classified either as a bulk sample or an eDNA sample depending on the concentration of the target DNA in it. For instance, 1 L of water collected from a river is an eDNA sample if the target taxon is fishes and a bulk sample if the target taxon is diatoms. Similarly, a Malaise trap catch is a bulk sample if the target group is insects but an eDNA sample if the vertebrates with which the insects have had interactions shortly before being collected is the group of interest. Failing to recognize this distinction can lead to major problems. For instance, applying a protocol developed for bulk samples on eDNA samples can render the sequencing results useless. But this is just one of the factors that can affect a metabarcoding study. Other factors include experimental design, laboratory processing and bioinformatic analysis (Alberdi, Aizpurua, Gilbert, & Bohmann, 2017).

Marker choice

One of the first and most important factors determining the results of metabarcoding is the choice of genetic marker that is going to act as barcode. A ‘marker’ is a gene fragment that is amplified using a certain primer pair. The ideal metabarcoding marker should consist of a short region (200-400 bp long) variable enough to discriminate between closely related species but not too variable between individuals of the same species, and flanked by conserved short regions where PCR primers with broad taxonomic coverage can attach (Deagle, Jarman, Coissac, Pompanon, & Taberlet, 2014). There is thus a compromise between discriminatory capacity and PCR amplification with ‘universal’ primers, as the former requires high variation and the latter requires low variation. The length of the marker is limited by the capacity of the sequencing platform. Currently, Illumina MiSeq and HiSeq are the most used systems for metabarcoding, and they are limited to short reads. In recent years we have started seeing use of other platforms with longer read lengths (Heeger et al., 2018), and it is possible that these will take over in the coming years.

For insects, as for other animals, the Folmer region of COI and shorter fragments (‘mini-barcodes’) within it (Hajibabaei et al., 2006; Meusnier et al., 2008; Yeo, Srivathsan, & Meier, 2020) were initially adopted as the standard metabarcoding markers. COI is a protein-coding mitochondrial sequence characterized by high sequence variability, especially in the third codon positions, giving it good discriminatory power. It is also relatively easy to amplify COI fragments, as mitochondria are abundant there are stretches where the COI sequence is highly conserved, at least at the first and second codon positions. Because the Folmer region was the original DNA barcoding marker, we now also have an extensive reference database for it (Andújar, Arribas, Yu, Vogler, & Emerson, 2018). However, two aspects make it difficult to use COI for metabarcoding of insects. First, the vast diversity of insects means that even the most conserved regions of the gene present some degree of variation at the amino acid level across the species one would want to amplify. Second, the redundancy of the genetic code means that the nucleotide base pair at the third position and sometimes also the first position of the codon can change without changing the corresponding amino acid sequence (Deagle et al., 2014). Given the formidable taxonomic diversity of insects, we are eventually likely to see most of these synonymous variants at the DNA level, and they will affect amplification success if they occur in the primer-matching regions of the sequence.

In the last decade, a wide array of ‘universal’ COI primers for insects have been designed, but most of them with very low or null degeneracy (Brandon-Mong et al., 2015). A degenerate primer is a set of alternative primer sequences that vary between two or more nucleotides at one or more positions. Primers with low degeneracy targeting variable regions amplify some species or higher taxonomic groups better than others, which can lead to an underrepresentation of the latter or the failure to detect them altogether. Such amplification biases have been documented for many primer pairs (Brandon-Mong et al., 2015; Hajibabaei, Shokralla, Zhou, Singer, & Baird, 2011; Morinière et al., 2016; Yu et al., 2012). In recent years, COI primers with high degeneracy have been designed to overcome this limitation, with very satisfactory results (Clarke, Soubrier, Weyrich, & Cooper, 2014; Elbrecht et al., 2019; Elbrecht & Leese, 2017). However, highly degenerate primers come with other problems, the most troubling of which may be the unspecific amplification of other organisms in the samples.

Another alternative is to abandon COI in favour of other markers associated with less serious amplification biases. These correspond generally to more conserved markers, such as the nuclear rRNA genes (18S is the most common) or the mitochondrial rRNA gene 16S. The 16S gene is the one that has received most attention. It is already a common marker in metabarcoding of vertebrates, especially mammals (Ji et al., 2020; Lynggaard et al., 2019), and it is more conserved than COI but less so than 18S. This facilitates primer design while still allowing discrimination between species (Clarke et al., 2014). The 16S gene has been tested as an alternative to COI for metabarcoding of insects, or as a complementary marker to be sequenced simultaneously (Clarke et al., 2014; Elbrecht et al., 2016; Epp et al., 2012; Kaunisto, Roslin, Sääksjärvi, & Vesterinen, 2017). It has been reported that the discriminatory capacity and the taxonomic coverage of 16S are too low for successful metabarcoding (Alberdi et al., 2017), but these problems could potentially be overcome by using appropriately degenerate primers and by sequencing longer fragments of the gene.

Effect of substrate for DNA extraction

As mentioned earlier, there are many factors influencing the results of a metabarcoding survey apart from marker choice. They include aspects of the sampling, laboratory processing and bioinformatic analysis (Deiner et al., 2017; Liu, Clarke, Baker, Jordan, & Burridge, 2019; Piper et al., 2019). One of the most important factors is the substrate used for the extraction of DNA. For instance, a metabarcoding analysis of insect diversity in soil or water samples (eDNA) will require different protocols and is likely to generate different results than an analysis based on the tissue of the insects caught in a trapping device, such as a Malaise trap (a bulk sample).

Malaise traps have been one of the most widely used devices for collecting insects since their invention in the 1930's (Malaise, 1937). Nowadays, it is common for ambitious national and global biomonitoring projects targeting insect faunas to rely largely or entirely on Malaise traps (Karlsson, Hartop, Forshage, Jaschhof, & Ronquist, 2020; see a global collecting and documenting effort at Global Malaise Trap Project: https://biodiversitygenomics.net/projects/gmp/). But the development and testing of metabarcoding methods for Malaise trap catches have been lagging behind that for other types of bulk samples, like samples of freshwater benthos or zooplankton (Blackman et al., 2019; Bucklin, Lindeque, Rodríguez-Ezpeleta, Albaina, & Lehtiniemi, 2016; Gibson et al., 2015; Shokralla et al., 2015) or eDNA samples from water and soil (Deiner et al., 2018; Dopheide, Xie, Buckley, Drummond, & Newcomb, 2018). Only very recently, the number of studies focusing on methodological aspects of metabarcoding of Malaise trap catches and other bulk samples from terrestrial ecosystems has started to rise (Creedy, Ng, & Vogler, 2019; Krehenwinkel et al., 2018, 2017; Morinière et al., 2016; Wilson, Brandon-Mong, Gan, & Sing, 2019).

While it is true that many protocols for metabarcoding of bulk samples of freshwater invertebrates are likely to be applicable to Malaise trap catches, this cannot always be assumed to be the case. For instance, several studies have shown that metabarcoding of the DNA from the preservative ethanol in which bulk samples of freshwater arthropods are collected or stored can successfully replace analyses of homogenized samples (Erdozain et al., 2019; Hajibabaei, Spall, Shokralla, & van Konynenburg, 2012; Martins et al., 2019; Zizka, Leese, Peinert, & Geiger, 2019). Analysis of preservative ethanol has several advantages: it would significantly reduce the time needed to process the samples, and it would leave the specimens intact for further taxonomic work (or any other type of work). However, this might be one of the cases where protocols are not transferable from freshwater to terrestrial bulk samples, as insects from these two environments have quite different characteristics. The insects found in freshwater environments are usually soft-bodied adults or larvae, while the ones found in terrestrial ecosystems are more often adults (particularly in Malaise trap catches, which are dominated by flying forms) and generally more sclerotized. This may explain why DNA extractions from ethanol of terrestrial samples have been found to underperform compared to tissue-extractions in species detection (Linard, Arribas, Andújar, Crampton-Platt, & Vogler, 2016).

Collecting traces of DNA in the environment (eDNA) is a non-invasive method of sampling the diversity of a taxon in a certain habitat. In the case of terrestrial insects, the best alternative would probably be soil samples, although airborne eDNA is an option worth considering, at least for flying or airborne species (Kraaijeveld et al., 2015). Soil eDNA would seem ideal for detecting insects that crawl or dig in or on leaf litter or the upper layers of soil, but it would also be likely to contain traces of insects in many other microhabitats that die and fall to the ground. Many flying insects have larval or pupal stages that are found on or in leaf litter or soil, so these species should also be present in soil eDNA.

However, analysis of eDNA comes with its own set of challenges. First, eDNA is usually found as extracellular DNA degraded into short fragments, so a marker that is suitable for bulk samples might not work for eDNA if it is too long. Second, insect DNA in environmental samples would probably be found in many cases in lower concentration than the DNA of other organisms, such as fungi, bacteria and nematodes. This can result in significant amplification of DNA from non-target taxa if the primer pair used is not specific enough for the group of interest. In the worst case, the sequenced samples can be completely dominated by off-target DNA, rendering them unsuitable for analysis of the diversity of the target group. Previous studies have shown off-target amplification to be a problem for COI analyses of soil eDNA using degenerate primers (Collins et al., 2019; Macher et al., 2018). Off-target amplification can also be a problem with less degenerate primers when the marker is more conserved, such as 18S (Yang et al., 2014). To avoid the problem, some researchers have used different markers for bulk samples and eDNA (Horton, Kershner, & Blackwood, 2017; Yang et al., 2014). However, this complicates the analysis, as the MOTUs (Molecular Taxonomic Units, a proxy for species based on similarity-clustering of sequences) obtained in the different types of samples cannot be compared without extensive sequence libraries providing cross-references between the markers. A completely different type of problem that may plague eDNA analysis is that certain insects present in the environment simply may not release enough DNA to the substrate to be detected. DNA is a stable molecule, so the reverse may also be true, that is, that DNA traces remain in the environment long after a species has disappeared.

The balance between morphological preservation and metabarcoding performance

If analysis of preservative ethanol is not satisfactory, what other options are there for metabarcoding of bulk samples that would leave the specimens intact for subsequent morphological work or additional genetic analyses? Non-destructive DNA extraction by temporary incubation of the insects in a mild lysis buffer might be a good option. If the treatment is mild enough, it should leave the insects in good condition, while potentially providing high-quality DNA extract for metabarcoding characterization of the sample. Another approach, which has been suggested, is to collect one leg of every individual in the sample and then homogenize and analyze all the legs in the sample (Beng et al., 2016; Yinqiu Ji et al., 2013). Although this preserves the specimens (except for the removed leg), it has the great disadvantage that it requires a vast amount of manual labour, so it may be out of reach for many projects because of time or resource constraints. Thus, mild lysis represents one of the most promising alternatives for metabarcoding projects that need to preserve the material for subsequent examination.

Although non-destructive extraction protocols have been used in some metabarcoding studies already (Yinqiu Ji et al., 2020; Vesterinen et al., 2016), there have been few tests of the performance of these protocols. In fact, only two studies so far have focused on methodological proofing of non-destructive DNA extraction. One of them used real samples and mock communities composed of macroinvertebrates from freshwater habitats (mainly insects, but also mollusks and annelids), and digested these samples in a lysis buffer from a commercial kit (Carew, Coleman, & Hoffmann, 2018). However, the results of this study should be considered with caution, for it has been mentioned that freshwater and terrestrial arthropods have different characteristics that can affect the outcome. The second study used mock communities composed of individuals collected in Malaise traps, and focused on comparing the results of metabarcoding when the complexity of the sample increases or the volume of the subsample of lysate used for extraction is reduced (Nielsen, Gilbert, Pape, & Bohmann, 2019). Neither of these studies compared the performance of different mild lysis protocols. Thus, it would be advantageous to count with more methodological work on mild lysis protocols before they are broadly implemented by entomologists.

Even if DNA extraction from preservative ethanol or from lysate aliquotes after mild digestion resulted in perfect metabarcoding recovery of the sample composition, there is still another conflict between molecular analysis and morphological study of the samples. Traditionally, insects have been preserved in 70 % ethanol for morphological examination (Martin, 1977). DNA extraction and PCR are known to work with specimens stored for a short period of time in 70% ethanol, but it has been shown that, in the longer term, the degradation of DNA is problematic for such applications (Baird, Pascoe, Zhou, & Hajibabaei, 2011; Carew, Metzeling, St Clair, & Hoffmann, 2017). Preserving the insects in high-grade ethanol (95-99 %) would seem to be the simplest solution, as ethanol in these concentrations is known to preserve DNA well, but it has not been completely implemented because of the received wisdom that high concentrations of ethanol make insects brittle and difficult to work with. To our knowledge, however, there has only been one study to date that has tried to address this phenomenon in a systematic way, and it only tested the effect of different concentrations of ethanol on the morphological preservation of three species of insects (King & Porter, 2004). Thus, there is a need for more extensive studies of this potential trade-off between preservation of insects for morphological study and for genetic characterization.

Objectives of the Thesis

The main objective of my thesis was to optimize metabarcoding of terrestrial insect community samples, so that this method can be used as an effective tool by insect biodiversity researchers in obtaining fast and accurate characterizations of the insect communities of interesting habitats, and in pinpointing potentially interesting species or specimens for further genetic or morphological analyses.
More specifically, I focused on:

- Design and evaluation of existing and new primers for all potential mitochondrial markers, not only COI, to minimize primer bias and to explore the potential of multi-marker approaches (**papers I, II**).

- Evaluation of different substrates for DNA extraction and insect community characterization, with a special focus on non-destructive alternatives to homogenization of Malaise trap samples (**paper II**).
- Optimization of preservation and processing protocols for maximizing metabarcoding performance while maintaining adequate morphological integrity of insect specimens (**papers III, IV**).

MATERIALS AND METHODS

In silico design and evaluation of primers and markers

For the design of the new primers targeting mitochondrial genes a dataset comprising all available mitochondrial genomes of Hexapoda from GenBank was downloaded in October 2015, while the evaluation of the newly designed primers as well as primers found in the literature was done over a second dataset downloaded in September 2016. The first dataset (D1) comprised 1,138 genomes belonging to 801 species, while the second dataset (D2) contained 1,600 mitogenomes from 1,081 species. This design allowed us to test the primers on species that were not present at the time of design, simulating what would happen in a real situation.

The design of the new primers was done following two pipelines, one based on the program `ecoPrimers` (Riaz et al., 2011), and the second using the software `DegePrime` (Hugerth et al., 2014). For the first pipeline, D1 was transformed into an `ecoPCR` database (Boyer et al., 2016) and the primers were designed with a length of 18 bp, zero mismatches allowed at the three last positions at the 3' end and an amplicon size ranging from 50 to 500 bp. The rest of the parameters were left at the default settings. For the second pipeline, all genomes in D1 were split to obtain the protein-coding and the rRNA genes using `Geneious` v8.1.7 (Kearse et al., 2012), and subsequently aligned using `MAFFT` v7.266 (Katoh & Standley, 2013). Primers of length of 18 bp were designed for each gene independently, for an amplicon size of 50 to 500 bp and two set-ups of maximum degeneracy: 12-fold and 216-fold. Once the primers were designed, the dataset D2 was transformed into an `ecoPCR` database and an *in silico* PCR amplification was done with `ecoPCR` (Ficetola et al., 2010) for the newly designed primers and primers from the literature. For the *in silico* PCR, no mismatches were allowed between primer and template, and only amplicons with a size ± 10 % of the expected length were permitted.

For evaluation, two properties were considered: taxonomic coverage – how much of the dataset is amplified with a given primer pair – and resolution capability – how well the resulting marker distinguishes between species. For measuring these two aspects, two indices have previously been proposed. Taxonomic coverage (B_C) measures amplification capacity as the proportion of species the primer pair can amplify from, and taxonomic resolution (B_S) measures resolution capability as the proportion of species that are unambiguously identified by the resulting marker among the species that are successfully amplified. However, in the definition of B_S, the oversplitting of MOTUs (the classification as two different "species" of sequences that belong to the same species) is not penalized. This could lead to an overestimation of the quality of a marker. To address this shortcoming, we proposed the index exclusive taxonomic resolution (B_E), which excludes from the group of 'unambiguously identified' those clusters that share the same species identity. Also, as B_E is defined as a fraction of B_C, we proposed the index effective taxonomic resolution (*ETR*), which is defined as the proportion of species unambiguously identified from the total present in the dataset. Lastly, to evaluate the performance of the use of several markers over the same dataset, we explored two approaches for combined primer design. In the first, *simultaneous combination*, we calculated the *ETR* of the two markers when the primers were designed using the entire dataset, while in the second, *residual combination*, we designed primers for a second marker using only the sequences of the species that were not amplified or correctly identified by the first marker.

The calculation of the B_C index was done with the script `ecotaxstat`, and the calculation of B_S with the script `ecotaxspecificity`, both from the `OBITools` package. The calculation of B_E was done using a custom pipeline using the algorithm `UCLUST` from the software `USEARCH` (Edgar, 2010).

Metabarcoding of eDNA and bulk samples

Samples were taken from three different points in the Nacka Nature Reserve, in the surroundings of Stockholm, at four time points during Summer-Autumn 2016. The three locations were chosen to maximize habitat diversity. At each location a Malaise trap was set up and run for a week. The insects were collected in 95 % ethanol, and unique bottles were used for each sample. In addition, a soil sample was collected at each trap location at the same time points. Each sample consisted of three replicates of soil cores within a radius of 20 m from the Malaise trap. Malaise trap and soil samples were stored at -20 °C for approximately six months before analysis. Immediately before analysis, soil samples were separated in leaf litter and humus (the first 2 cm of the soil core) and homogenized with a mortar after ultrafreezing in a bath of liquid nitrogen. Once homogenized, the three replicates of each sample were pooled together and mixed. From each sample 0.4 g of soil or leaf litter were extracted with the Nucleospin Soil kit (Macherey-Nagel, Germany). With respect to Malaise trap catches, the ethanol of each bottle was passed through a 0.6 mm sieve to retain small animals or body parts, and filtered with a vacuum pump using 0.45 µm Durapore membrane filters (Merk, Germany). The filter was folded and stored in the lysis buffer of the KingFisher Cell & Tissue kit (Thermo Fisher Scientific) in a 2 mL tube and frozen until DNA extraction. The insects in the sample were dried on filter paper and then homogenized in a mortar after quick immersion of the mortar in a bath of liquid nitrogen. The resulting slurry was then stored in tubes in the lysis buffer of the KingFisher Cell & Tissue kit and frozen. For DNA extraction both filters and tissue slurry were incubated overnight at 56 °C and subsequently DNA was extracted using the KingFisher Cell & Tissue kit on a KingFisher Duo extraction robot (Thermo Fisher Scientific, USA).

A fragment of 322 bp of COI was amplified from each sample with the primers BF2-BR1 (Elbrecht & Leese, 2017), and a fragment of ~345 bp of 16S was amplified using the primers Chiar16SF-Chiar16SR (**paper I**). These primers were selected as the best-performing primers in **paper I**. Both primer pairs had attached at the 5' end a unique 8 bp long tag for sample multiplexing (Binladen et al., 2007). The PCRs were carried out with Illustra Hot Start Mix RTG beads (GE Healthcare Life Sciences), run in duplicates and then pooled together before library preparation. Libraries were prepared using the TueSeq PCR-free kit (Illumina, USA), consisting of enzymatic ligation of the adapters to the amplicons, and sequenced on a Illumina MiSeq using the v3 chemistry 2x300 PE reads at SciLifeLab (Stockholm).

Non-destructive lysis and DNA extraction methods

For the evaluation of non-destructive extraction protocols, a set of ten types of mock communities was prepared, each of them with four replicates (40 tubes in total). A total of 23 species were used for the communities, obtained from donations of standardized cultures from other laboratories and from the NRM dermestarium, commercially purchased or personally collected. The communities contained only 22 species, meaning that in each community type there was a species missing that was present in the others. All individuals of a species were selected to be of approximately the same size, and a few representatives were weighed to obtain an estimate of the average biomass of the selected specimens. An additional individual from each species was used to generate a barcode reference library for COI and 16S. The COI barcode was amplified using the primers jgLCO1490-jgHCO2198 (Geller, Meyer, Parker, & Hawk, 2013), except for *Formica rufa*, which was amplified using LepF1-LepR1 (Hebert et al., 2004). The 16S barcode was amplified using the primers 16Sar-16Sb2 (Cognato & Vogler, 2001; Simon et al., 1994). The PCRs were conducted using Illustra Hot Start Mix RTG beads and Sanger-sequenced at Macrogen Europe B.V. (Amsterdam, Netherlands).

The mock communities were then digested under four lysis treatments by combining two digestion buffers and two incubation times. The first buffer (B1) was moderately aggressive, containing the basic compounds (EDTA, SDS, NaCl, Tris-HCl) and 0.1 % by volume of proteinase K (Aljanabi & Martinez, 1997; Vesterinen et al., 2016). The second buffer (B2) was more chemically aggressive, and slightly different in composition ($CaCl_2$, SDS, NaCl, Tris-HCl), with the addition of dithiothreitol (DTT) and 1 % by volume of proteinase K. The two incubation times were 2.5 (LT1) and 5 (LT2) hours. After incubation, the lysate (the lysis buffer containing the DNA and digested tissue) was decanted before proceeding with DNA extraction. The insects were then first rinsed with distilled water, rinsed again with 70 % ethanol and finally stored in 80 % ethanol. The lysate of each tube was purified using two methods of extracting the DNA. The two purification methods were a manual protocol involving precipitation of proteins with a saturated salt solution followed by precipitation of the DNA with isopropanol (P1) (Aljanabi & Martinez, 1997), and an automated protocol using a laboratory robot with a commercial kit (P2). An important difference between the two methods is that purification P1 used 7.5 mL of lysate as starting volume and purification P2 had an input volume of 225 μL. The DNA extraction from all the lysates using both purification methods generated a total of 80 samples (10 mock communities x 2 lysis buffers x 2 incubation times x 2 purification methods).

Once extracted, the concentration and ratio of absorbance at 260 nm and 280 nm (a measure of the proportion of DNA to proteins in the extract; values between 1.8 and 2.0 are considered optimal, being 2.0 the ratio of a pure DNA solution) were measured using a NanoVue instrument (version 4282 v1.7.3, GE Healthcare Life Sciences). The same fragments as in **paper II** were amplified using the same primers. However, in this case the library preparation was done following the two-step PCR protocol. This protocol consists of a first PCR, in which the 5’ ends of the marker-specific primers are attached to part of the Illumina adapter. Then, a second PCR is done with the complete adapter as the primer, so that, in the end, all amplicons have the entire Illumina adapter attached to both ends. The libraries were sequenced on a Illumina MiSeq using the v3 chemistry 2x300 PE reads at SciLifeLab.

Assessment of ethanol-induced brittleness on insects

To examine the effect of high concentrations of ethanol on the fragility of insects, mock communities containing seven species (spanning four orders) were used. The number of individuals of each species varied between ten and two depending on the size of the species and their availability in large quantities. The insects were obtained alive from commercial providers, standardized cultures or manually collected, and killed by either freezing them or submerging them in ethanol. The mock communities were kept in increasing concentrations of ethanol, from 30 to 99 %, for a month, after which three experiments were conducted.

The first experiment consisted of subjecting the insects to two shaking regimes: a gentle one in which they were vortexed for a minute, and a more vigorous one in which the vortexing time was doubled to two minutes. In the second and third experiments, only two concentrations of ethanol were used: 70 % (standard for morphological preservation) and 95 % (standard for DNA preservation). For the second experiment, the tubes were carried in the backpack of two experimenters. One had to walk cautiously without sudden moves, while the other was requested to run at suitable occasions. Also, two parcels containing the experimental tubes were mailed using the Swedish national post service. Finally, for the last experiment, the insects were subjected to the gentle treatment of experiment one, but they were previously treated by drying or repeated freeze-thaw cycles. Specimens left in tubes without such pretreatments were used as controls.

Upon completion of the experimental treatments, all individuals were inspected for loss of appendages using a stereomicroscope. The appendages examined were legs, wings, antennae and head (with variations for each species, *e.g.* excluding wings for the ants). The number and type of appendages lost were recorded and compared. Statistical analyses of the data were conducted using R v3.3.3 (R Core Team, 2017). A generalized linear mixed-effects model, with the loss of appendages as a function of treatment and concentration, and replicate as a random effect, was fitted to the data using the package 'glmmTMB' (Brooks et al., 2017). To analyze the effects of ethanol concentration and experimental treatment, an analysis of variance (ANOVA) was conducted, followed by a Tukey test to analyze pairwise differences between means using the package 'emmeans' (Lenth, Singmann, Love, & Others, 2018).

Bioinformatic and statistical analysis of the metabarcoding data

The bioinformatic processing of the sequencing data was conducted using a pipeline based on `OBITools` in combination with functions from other programs, such as `CUTADAPT` v1.8.0 (M. Martin, 2011), `VSEARCH` (Rognes, Flouri, Nichols, Quince, & Mahé, 2016), `SWARM` (Mahé, Rognes, Quince, de Vargas, & Dunthorn, 2015), `LULU` (Frøslev et al., 2017) and scripts from the Metabarpark GitHub repository (https://github.com/metabarpark/R_scripts_metabarpark), as well as custom scripts. In short, reads were quality checked and pair-end merged. Demultiplexing was done immediately after or before merging, depending on the library preparation method, and only sequences of the expected length were kept. Chimeras were removed based on sequence similarity and abundance, and reads that passed the filter were clustered into MOTUs with flexible clustering thresholds of 3–4 % for COI and 1–2 % for 16S. Small MOTU clusters with high sequence similarity to larger ones were included in the latter when the clusters showed patterns of co-occurrence. The most abundant sequence of each MOTU was kept for taxonomic annotation by comparison with a custom reference database obtained either from BOLD, from GenBank and EMBL, or by Sanger-sequencing the species included in the mock communities. The final dataset was refined by aggregation of MOTUs with coincident species-level identification and removing MOTUs with less than ten reads in total.

All statistical analyses were done in R v3.3.3. For the analysis of the recovered communities from different substrates (**paper II**), a nonmetric multidimensional scaling (NMDS) based on a Jaccard dissimilarity matrix (which considers only presence/absence) was used to visualize the differences between the species detected in the filtered ethanol and those detected in the tissue homogenate. A permutational analysis of variance (PERMANOVA) was run to determine whether there were significant differences in the recovered community from each trap sample. This was done using the package ‘vegan’ (Oksanen et al., 2013). To determine whether size or degree of sclerotization determined the probability of detection of an insect family in either of the two substrates, we fitted a generalized linear mixed effects model with the detection as a function of sample type, size, sclerotization and their two- and three-way interactions as fixed effects, and with sample ID as random effect.

For the analysis of the data from the non-destructive extraction methods (**paper IV**), we used a split-split-plot design, with buffer type as main plot, incubation time as subplot, purification as sub-subplot, and community type (A-J) as replicate. We then used analysis of variance (ANOVA) to determine whether any of these factors had a significant effect on the concentration and purity of the DNA extracts, using the package ‘agricolae’ (de Mendiburu, 2014). Another ANOVA was used to determine if the different lysis treatments and purification methods had a significant effect on the number of species recovered by each marker. In addition, we used regression analysis to examine the relationship between relative read abundances and relative abundances in the artificial communities measured by the number of individuals or biomass. Finally, we assessed the difference between estimates of the community composition based on the two metabarcoding markers and the true composition in terms of the number of individuals or biomass. First, we compared the estimated alpha diversity of each sample, assessed using the Shannon index (H’) computed using the package ‘vegan’, with that of the corresponding true community. Second, we measured the Kullback-Liebler divergences between the metabarcoding estimates of the community composition and the true composition using the package ‘LaplacesDemon’ (Statisticat, 2018). Then, an ANOVA was used to determine whether there were significant differences in these measures based on buffer, lysis time and purification method.

RESULTS AND DISCUSSION

Paper I

The genes ATP8, ND2 and ND6 were challenging to extract bioinformatically from the published mitochondrial genomes because of considerable structural variation in them across insects. This variation would also make it difficult to use them for metabarcoding, so they were excluded from further analyses. For the remaining genes (12S, 16S, COI, COII, COIII, ND1, ND3, ND4, ND4L and ND6), only those primers matching at least half of the sequences in D1 were considered for the rest of the analyses. With `DegePrime`, primers fulfilling this requirement were found for ATP6, ND1, ND3 and ND4 only when maximum degeneracy was set to 216-fold, and none were found for ND4L with either 12- or 216-fold degeneracy. With `ecoPrimers`, which does not accommodate degeneracy, only five pairs of primers meeting this criterion were found, all targeting regions in the 16S rRNA gene.

Using the primers with highest coverage for each gene among those designed with `DegePrime`, we observed the expected increase in the value of B_S with increasing similarity threshold. The maximum value was not reached until the similarity threshold was set to 100 %, that is, when only identical sequences were considered to belong to the same species (Figure 1). In contrast, the value of B_E decreased brusquely as the threshold approached 100 % similarity. The maximum B_E values were obtained at different similarity thresholds for each gene, proportional to

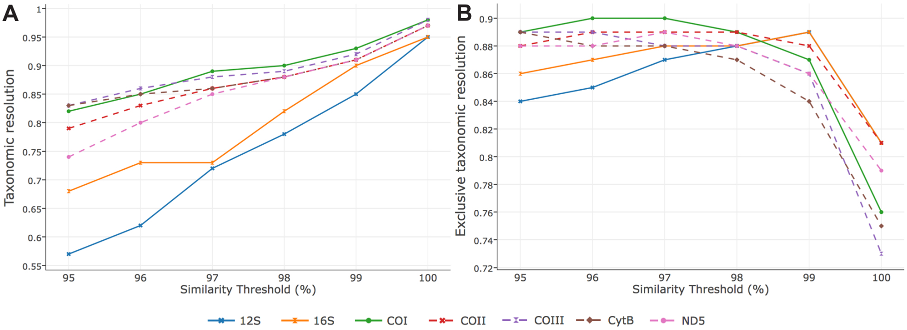


***Figure 1** Taxonomic resolution (left) and exclusive taxonomic resolution (right) at different similarity thresholds. For all genes, the B_S index increases monotonously as the similarity threshold increases, while B_E peaks at different points, indicating where the barcoding gap is for each marker.*

the rate of sequence evolution of that gene. These results show that B_S is not a reliable index to measure taxonomic resolution, as it fails to discriminate between intra- and interspecific genetic diversity. This might not be a problem in an ideal situation, when a perfectly complete reference database for the employed marker is available. However, this is almost never the case, even with COI. Consequently, the downstream analyses were done using B_E as the measure of taxonomic resolution.

We found that `DegePrime`-designed primers clearly outperformed those designed using `ecoPrimers`, mainly because of the lack of degeneracy in the latter. The primers designed with `ecoPrimers` had a maximum value of B_C of 0.6, and average values of B_E of 0.8. This low value of resolution is due to their short length, as the sizes ranged from 80 to 140 bp. The `DegePrime`-designed primers also clearly outperformed the primers from the literature that do not include any degenerate positions. Increasing the degeneracy, even to only 12-fold, had a very positive effect on the coverage of the primers tested, particularly for the more conserved genes (the rRNA genes 12S and 16S). Published primers for 16S with degeneracy from 2 to 12 perform better as degeneracy increases, but their shorter size compared to the ones designed with `DegePrime` do not allow them to reach high values of B_E. To reach values of B_C comparable to the rRNA genes, primers for protein-coding genes required a much higher degeneracy (192- or 216-fold), while for 12S and 16S, an increase in degeneracy from 12-fold to 216-fold did not translate into a significant difference in B_C. Published primers for COI with this level of degeneracy performed as well or even better than the primers designed with `DegePrime`. But `DegePrime`-designed markers from other genes like COII, COIII and CytB with high values of B_C also pro-

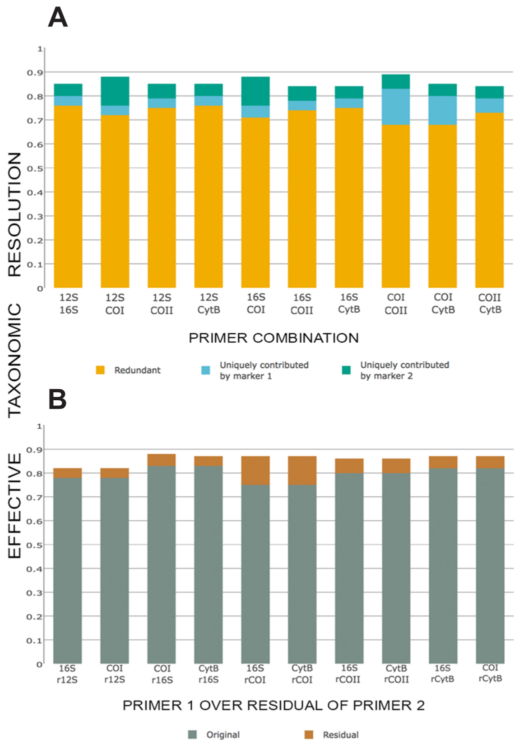


Figure 2 *Combined effective taxonomic resolution (ETR) for all the markers with an ETR ≥0.75. Top: simultaneously combined ETR showing redundant ETR (orange), uniquely contributed ETR by the first primer pair (blue) and uniquely contributed ETR by the second primer pair (green). Bottom: residually combined ETR showing original ETR (grey) and residual ETR (brown).*

vided similarly high values of B_E. Separate analysis of the markers for the most diverse insect orders and those most commonly found in traps showed that the 16S marker presented the least amplification bias, only paralleled by COI primers with very high amplification (and high levels of degeneracy). This is consistent with results from previous studies (Brandon-Mong et al., 2015; Clarke et al., 2014; Elbrecht et al., 2016). However, the scarce reference data for other markers than COI make it more difficult to use them for metabarcoding, and the use of other markers than COI is therefore opposed by some authors (Andújar et al., 2018). The situation is somewhat better for 16S than for the other markers, but even the 16S reference libraries are far behind those for COI.

We found that the combination of two markers could increase the value of *ETR* with any of the two combination approaches. The best combinations were those involving COI, and more specifically, the simultaneous combination between COI and 16S (Figure 2). In the residual approach, those primers targeting the gene that acted as original marker, that is, primers for 16S that were designed using the species not amplified by the first 16S primers, showed the lowest ETR_C values. That is, it was always better to choose another gene for the second marker. Such a multilocus metabarcoding approach has been used previously in other studies targeting not only insects, but also other diverse groups or studies with a broad taxonomic scope (Cowart et al., 2015; Drummond et al., 2015; Shaw et al., 2016). It was shown that one marker can fill up the gaps for those taxonomic groups not amplified or properly resolved by the other one (Wangensteen, Palacín, Guardiola, & Turon, 2018). The fact that for insects this approach is less popular might be due to the use in a previous study using multilocus barcoding of 16S primers with poor performance (Alberdi et al., 2017).

Paper II

In total, between the trap catches and the soil samples, we detected 432 MOTUs with COI and 430 with 16S. Nevertheless, the number of MOTUs in each substrate greatly varied between the two markers, as well as the distribution of MOTUs in different taxonomic groups. With COI, most MOTUs were detected in the substrates from the traps (ethanol and tissue) and only 14 of the 432 MOTUs were detected in the soil samples. With 16S, the number of MOTUs found in the soil (leaf litter and humus) was 120, almost ten times more than with COI. Overlap between markers was also low in the trap samples. Almost as many arthropod families were detected only by one marker as those detected by both markers simultaneously (Figure 3). The analysis of taxonomic annotation overlap is dependent on the size of the reference databases. Considering that the reference database for COI was almost 10 times larger than the one for 16S, it is possible that the number of families recovered only by 16S was even higher. This demonstrates that the multilocus approach increases biodiversity detection not only in those studies dealing with a very wide taxonomic scope, *e.g.* zooplankton, but also in insect studies, which is in concordance with previously published results (Holman et al., 2019; Kaunisto et al., 2017; Thomsen & Sigsgaard, 2019; Wangensteen et al., 2018).

The advantage of using 16S is obvious when we look at the data from the eDNA samples and compare it with bulk samples. The proportion of reads assigned to arthropods with COI dropped from close to 100 % of the sample to 50–80 % and then to almost 0 % when going from tissue to ethanol and to soil substrates, while with 16S the decrease was from around 100 % for the tissue to 80–85 % for the rest of the substrates. This difference is likely caused by the high levels of degeneracy of the COI primers, resulting in amplification of many other taxonomic groups than those which are the target when significant amounts of foreign DNA is present. The primers used for 16S, although with similar or higher coverage than the highly degenerate COI primers, are only slightly degenerate and more specific for the target group. Excessive amplification of non-target groups can cause problems in environmental surveys using eDNA, as the target DNA is not dominant in the sample (Collins et al., 2019; Macher et al., 2018). Thus, for any ecological study using eDNA samples or combining them with bulk samples, it would be recommended to include 16S (or another, equally specific primer), to allow for direct comparisons between substrates.

Finally, we found that the community recovered was very different also between preservative ethanol and tissue homogenate from the same Malaise trap samples (Figure 4). Independently of the marker, the community recovered from the ethanol was not representative of that recovered from the tissue, and, more importantly, neither was the former a subset of the latter. We found that the detection in ethanol or tissue of

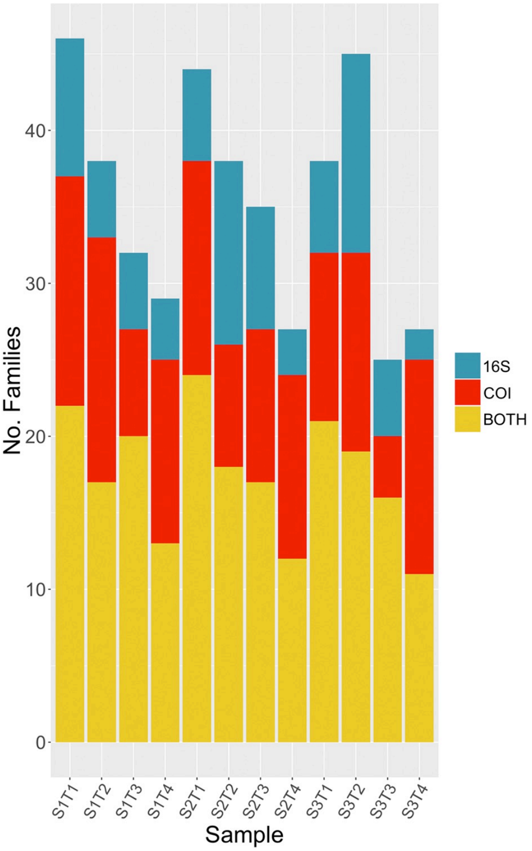


Figure 3 *Families recovered exclusively by 16S (blue) or COI (red) markers from each of the trap samples (i.e. bulk tissue and preservative fluid), or detected by both markers (orange).*

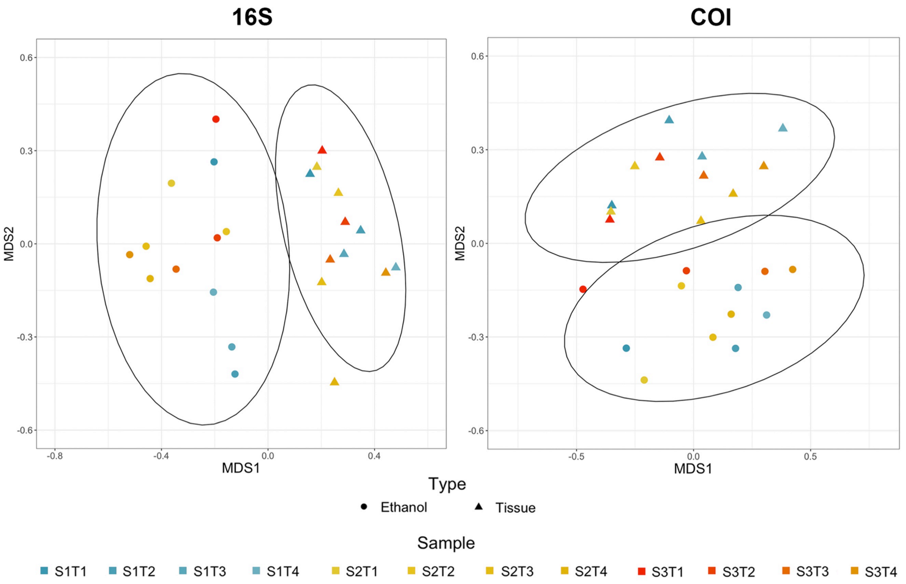


Figure 4 *NMDS plots, based on a Jackard dissimilarity matrix, of the recovered communities from ethanol and tissue for the 16S marker (left) and the COI marker (right).*

a given insect family was determined by the size and degree of sclerotization of its member species. Large and sclerotized insects were more likely to be detected from the tissue substrate, while the small and weakly sclerotized ones were preferentially recovered from the ethanol. In addition, for those families that were recovered both from tissue and ethanol, the sequence read abundance corresponding to each of the substrates followed the same pattern. Thus, despite the promising results of earlier metabarcoding studies using preservative ethanol from samples of freshwater fauna (Erdozain et al., 2019; Hajibabaei et al., 2012; Martins et al., 2019; Zizka et al., 2019), our results demonstrate that this may not be a good approach for terrestrial insect samples, as the presence of heavily sclerotized insects – rare in freshwater habitats – would probably be missed due to their DNA not being leaked through the body wall to the ethanol.

Paper III

All mild lysis treatments tested in **paper III** produced DNA extracts suitable for PCR, while maintaining the morphology of the insects in a good state of preservation. Not only the exoskeleton but also other morphological features, such as colouration, were maintained, demonstrating that effective DNA extraction for metabarcoding is compatible with good morphology preservation for bulk samples of terrestrial insects. The success was likely due to in part to the short periods of incubation (2.5 or 5 hours), more similar to those used in mild lysis protocols for freshwater invertebrate samples (Carew et al., 2018) than to those used previously for samples of terrestrial insects, spanning from 14 to 72 hours (Ji et al., 2020; Nielsen et al., 2019; Vesterinen et al., 2016).

The different lysis treatments and purification methods all had significant effects on the concentration and purity of the DNA extracts (Figure 5). Incubating the samples in buffer B2 led to higher concentrations of DNA, as did incubating the samples for a longer time (LT2). The manual purification that started with 7.5 mL always resulted in higher concentrations of DNA than the automated one. An interesting result is that the DNA concentration was not affected by the incubation time when purifying with the robot. The most probable explanation for this is that the amount of magnetic beads used in the robot was constant, binding the same amount of DNA regardless of the amount of DNA available in the lysate.

These overall results are expected, as a more chemically aggressive buffer, a longer digestion time and a larger input volue are all likely to increase the con-

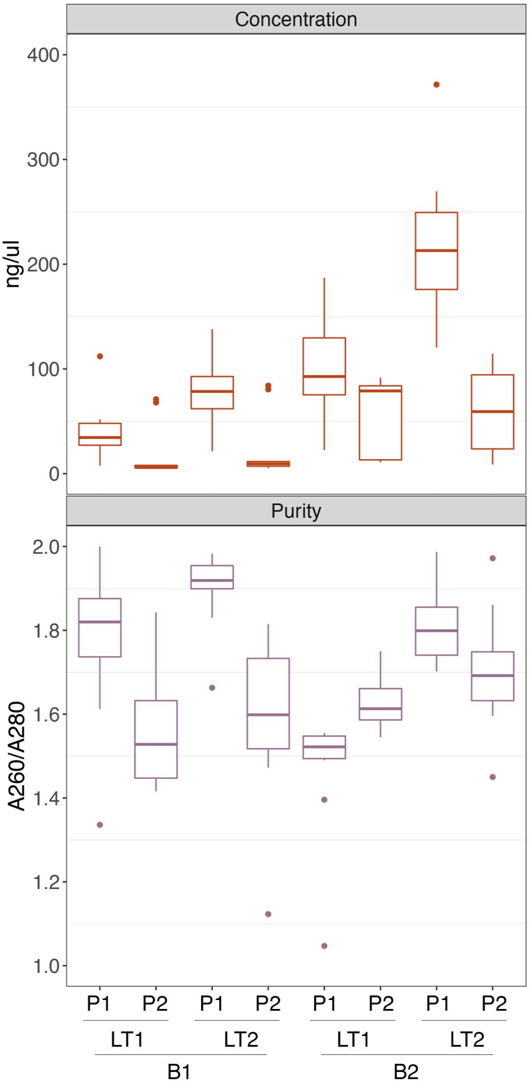


***Figure 5** Concentration and purity of the DNA extracts from different extraction methods. DNA concentration (upper panel) clearly increases with buffer aggressiveness and incubation time using the manual salt saturation purification protocol, while the increase due to incubation time is less clear, but the effect of lysis buffer can still be appreciated when using the automated robot purification protocol. Purity of the DNA extract (lower panel) is higher for the manual purification and the longer incubation times, regardless of the lysis buffer. B: buffer; LT: lysis time; P: purification method.*

centration of DNA in the extract. This contrasts with previous results in which the DNA concentration did not differ significantly between manual and automated purification (Nielsen et al., 2019), but in that previous case, the starting volume was similar for both types of purification. Regarding the purity of the DNA extracts, neither buffer type nor the three-way interaction between the buffer type, incubation time and purification method had a significant effect on the value of the ratio A260/A280. When applied to samples incubated with buffer B1, the manual purification method (P1) produced extracts of higher purity than the automated method (P2). This is consistent with the fact that manual protocols are usually recommended for difficult samples with inhibitors (for instance, mollusks, platyhelminths or nemertines that produce abundant mucus that hinders the extraction) because of their efficiency. For samples incubated with buffer B2, only the longer incubation time (LT2) increased the purity. A possible explanation for this is that the longer incubation in this buffer, which had a higher concentration of proteolytic compounds, allowed the enzymes to hydrolyze the proteins more effectively, thus increasing the purity. Although some A260/A280 values were low (down to 1.05 in some cases), the average purity was fairly good (1.5 to 1.9) and, as mentioned, enough to produce adequate PCR products.

The differences in concentration and purity of the DNA extract, however, did not translate to large differences in the number of species detected. For 16S, there were no significant effects of any of the three factors (lysis buffer, incubation time or purification method) on species recovery. However, for COI, both the buffer and its interaction with purification method had an impact. The combination B2–P1 provided the highest response in terms of species recovered, with B2–P2 second, then B1–P1 and finally B1–P2. However, although significant, these differences had only a small effect on species recovery. This differs from another study, in which the salt saturation method was shown to provide metabarcoding data with higher species richness than commercial kits did (Kaunisto et al., 2017). However, this study used faecal samples, while ours was based on fresh and well preserved bulk samples, which could explain the different outcomes. Our results are in line with a more recent study (Nielsen et al., 2019) that used mock insect community samples like our study, and that showed no differences in species recovery between the methods.

With respect to the accuracy in recovering the true species composition of the sample, including the relative abundances in terms of specimen numbers or biomass, we showed that non-destructive lysis methods can be optimized to represent the original sample more precisely. Samples incubated in the less chemically aggressive buffer (B1) produced metabarcoding estimates of the Shannon diversity index that were more similar to the Shannon diversity of the real mock communities (measured either in terms of individuals or biomass), than samples incubated in buffer B2. This was the case for both COI and 16S markers, and independent of the incubation time. However, for the Kullback-Leibler divergences between metabarcoding estimates of community composition and real community composition, the situation was different (Figure 6). For 16S, incubation in buffer B1 for a short time generated estimates with the lowest divergences with respect to the real communities, while incubating for a longer time or in buffer B2 increased the divergence. For COI, only the incubation in the more aggressive buffer produced samples with higher values of the divergence, but these values were strongly affected by sequencing depth. A possible explanation is that a greater sequencing depth skewed the distributions of the proportion of reads in the samples when they were transformed to a total of 1. These patterns can potentially be explained by the relation between body surface and body volume in combination with the length and chemical aggressiveness of the lysis (Nielsen et al., 2019). At the beginning, both large and small individuals will release DNA proportionally to their exposed surface (proportional to the square of the size), while as the incubation continues, individuals will start to release DNA from the internal tissue (proportional to the cube of the size). Thus, larger insects will contribute proportionally more to the DNA pool the longer the lysis period, or the more invasive the digestion buffer. We predict that, had the samples been homogenized, lower values of H' and higher values of the Kullback-Leibler divergence would have been observed, in line with the trends we observed for the tested mild lysis methods.

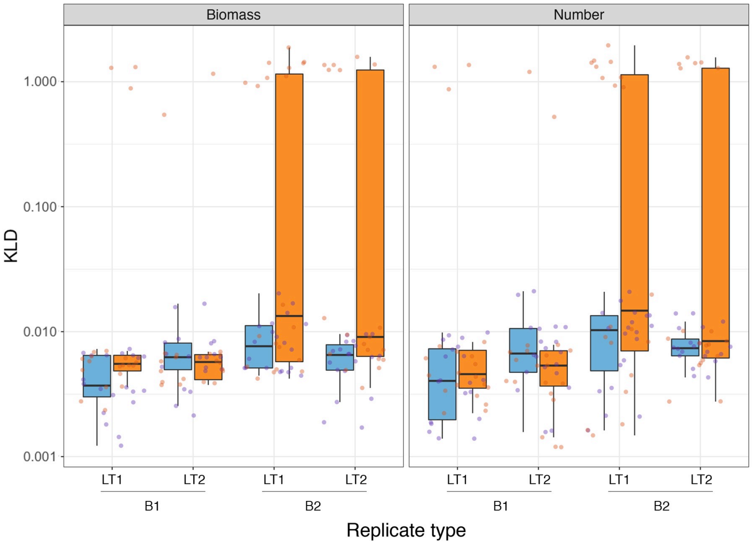


Figure 6 *Kullback-Leibler divergences between the true community composition and the metabarcoding estimates of it. Community composition is measured in terms of biomass (left) or the number of specimens (right). Data are shown both for the 16S marker (blue) and the COI marker (orange). For the 16S marker, the divergence between the metabarcoding and the original sample increases with buffer aggressiveness and incubation time, while for the COI marker, only an increase in buffer aggressiveness increases the divergence. Note that the y axis is on a logarithmic scale.*

Paper IV

In the study of the effects of different concentrations of ethanol on the preservation of insects (**paper IV**), we observed marked differences between treatments and concentrations, as well as between the experimental species. More strongly sclerotized or robust species (*Formica*, *Dermestes*, *Dacnusa*) were less susceptible to the fragility induced by ethanol. However, other species, like *Aphidoletes*, *Macrolophus* or *Calliphora* were more prone to lose appendages when stored in suboptimal concentrations of ethanol.

Our results show that, indeed, ethanol concentration has a significant effect on insect brittleness, assessed by appendage loss under different vortex shaking regimes, at least in many of the studied species (Figure 7). Intermediate concentrations of ethanol (70–80%) were usually associated with the least amount of appendage loss. In high concentrations of ethanol, most species lost more appendages. Interestingly, low concentrations (30–50 %) were also associated with increased fragility in some species.

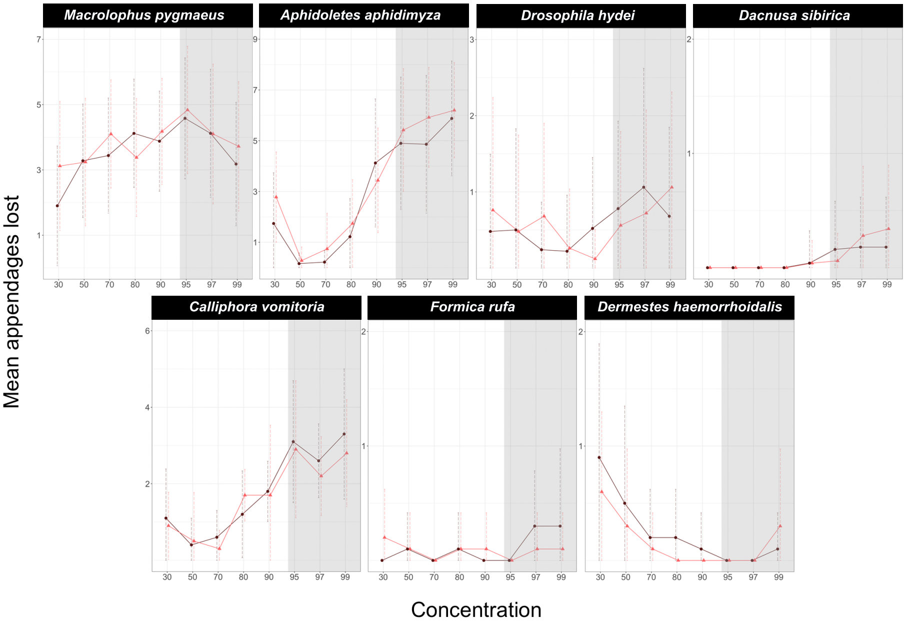


***Figure** 7 Effect of ethanol concentration on the number of appendages lost by different insect species. Dark purple circles represent the Gentle shaking regime while bright red triangles represent the Vigorous regime. The shadowed area corresponds to the ethanol concentrations in which DNA is optimally preserved according to literature.*

Our transport experiment indicates that a careless carrier (the experimenter that was encouraged to run to catch public transportation) can be more damaging than shipping the samples by regular mail. Finally, our results on drying and on freeze-thaw cycles failed to reveal any differences between ethanol concentrations; all the tested species seemed robust to these treatments. In fact, somewhat surprisingly, *Macrolophus* (Heteroptera, Miridae) lost fewer appendages after being subjected to cycles of freezing-thawing than in the control treatment. However, the specimens were significantly affected by this treatment; in particular, soft body parts like the abdomen were noticeably shriveled up.

Our results support the received wisdom that high concentrations of ethanol induce brittleness in insects. Even though the effect varies greatly among species depending on their characteristics, nearly all species are optimally preserved at intermediate ethanol concentrations (70 or 80 %). This means that there is a conflict between morphological and DNA preservation, as long term storage at these intermediate ethanol concentrations leads to noticeable degradation of the DNA (Baird et al., 2011; Bisanti, Ganassi, & Mandrioli, 2009; Mandrioli, 2008). Some alternatives have been proposed in the literature, such as storing the insects in 70 % after an initial period of storage in 95 % to preserve the DNA (Stein, White, Mazor, Miller, & Pilgrim, 2013). However, it is this first initial phase of rapid desiccation in high-grade ethanol that is supposedly most damaging for morphological preservation (Martin, 1977). Currently, there is no optimal storage protocol for preserving insects both for morphological and molecular study, and compromises must be made in one direction or the other in large-scale collecting campaigns that aim to make use of both sources of information.

CONCLUDING REMARKS

Metabarcoding can be a powerful tool at the service of taxonomists and ecologists. However, the method is still in its infancy, and some protocols are clearly better than others although we still lack sufficient data to provide firm recommendations. The goal of this thesis was to address precisely this problem by optimizing metabarcoding protocols for insect biodiversity studies. Although the target group was terrestrial insects, many of the results presented here are likely to be applicable also to metabarcoding of other eukaryotic organisms. Even if the results are not directly transferable, the optimization strategies can usually be generalized.

In **paper I** we developed a pipeline for designing and evaluating PCR primers and their corresponding markers. We proposed new indices for evaluating metabarcoding performance that we believe are more biologically accurate than previous ones, and we showcased a pipeline for primer design that has not been widely implemented in animal metabarcoding, despite its advantages. Also, we found that COI is not the best marker for metabarcoding of insects, despite its widespread use, but 16S, given the presence of very highly conserved windows suitable for primer design flanking a variable region and the reduced amplification biases. However, the lack of reference data is a major disadvantage in implementing 16S metabarcoding. As a compromise, we advocate for the combined use of both markers.

In **paper II** we used the two best performing primers from **paper I** to test the multilocus approach. This resulted in enhanced biodiversity detection compared to using only COI. In addition, the advantage of using the 16S primers, with much lower levels of degeneracy, was revealed after metabarcoding of the soil eDNA samples, in which the off-target amplification due to the high degeneracy of the COI primers produced poor results for the target group, arthropods. Also, we tested the suitability of using the preservative ethanol of the Malaise trap catches as substrate for DNA extraction for non-destructive metabarcoding of the insects in these samples. However, the communities recovered from these two substrates were significantly different, demonstrating that analysis of preservative ethanol is not a satisfactory replacement for analysis of tissue homogenate. Neither were the preservative ethanol results a subset of the tissue homogenate results, suggesting that there might be mild lysis protocols that could combine the advantages of these two methods. Our analysis indicated that the cause of the difference between preservative ethanol and tissue homogenate is that the leakage of DNA from the insects into the ethanol is strongly dependent on the degree of sclerotization of the different insect groups.

In **paper III**, we examined the possibility of using non-destructive lysis protocols in obtaining good-quality DNA for metabarcoding from bulk samples of insects, while preserving their taxonomically informative features intact. We found that even a mild lysis applied during a short time interval (2 h 30 minutes), and combined with DNA extraction performed with a commercial kit on an automated instrument, was able to generate DNA extracts of sufficient quality to retrieve an accurate representation of the samples with metabarcoding. Furthermore, we found that a mild and quick lysis resulted in the most accurate quantitative metabarcoding estimates of community composition. This protocol also generated diversity index estimates that were comparable to those of the actual mock communities analyzed.

Lastly, we exposed in **paper IV** a trade-off between preservation of DNA and preservation of morphology. Specifically, high ethanol concentrations required for optimal DNA preservation are detrimental for the preservation of morphology. This poses a problem for modern-day taxonomy, which is no longer restricted to the examination of morphology, but instead tends to combine morphological study with genomic analyses. The negative effect of high ethanol concentrations on morphological preservation is not universal. Some groups seem to be relatively robust to these effects, while others, such as Diptera (which represents a large fraction of most insect faunas of the world) are severely affected. Undoubtedly, there is still much to be learned about the preservative effects of different concentrations of ethanol. However, our studies show that these effects are important to consider in projects collecting material for both morphological study and genetic analysis, such as metabarcoding.

In the coming years we will witness the routine use of metabarcoding of Malaise trap samples in large-scale collecting campaigns (for ongoing or planned efforts here in Sweden, see the Insect Biome Atlas (https://www.insectbiomeatlas.com) or the LifePlan (https://www.helsinki.fi/en/projects/lifeplan) projects), the same way that metabarcoding is now starting to be applied to freshwater samples. In these studies, metabarcoding will be a cornerstone of sophisticated biodiversity data analyses. I hope that the work presented in this thesis will be useful for researchers working in this field. The fast progress in machine learning and automated image identification, coupled with metabarcoding, will surely represent a revolution similar to that of the introduction of metabarcoding itself. The analysis of intraspecific genetic variability from metabarcoding data (e.g. metaphylogeography) is another area where we are likely to see major break-throughs in the coming years, even though the analysis of intraspecific variability will prove challenging at the bioinformatic level. For these applications, it will be necessary not only to discriminate between species but also, more importantly, to discriminate between natural and artifactual variability among sequences within species.

I am very happy to have contributed to a field that not only taxonomists and ecologists can benefit from, but also those working in the interface between research and decision-making, as evidenced by the increasing number of companies and laboratories using metabarcoding and working hand in hand with environmental agencies and governmental institutions. As scientists, we all have the desire to have a real impact on society and, as biologists, having our work helping to better take care of the planet is a real privilege.

REFERENCES

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2017). Scrutinizing key steps for reliable metabarcoding of environmental samples. Methods in Ecology and Evolution, 17, 730–714.

Aljanabi, S. M., & Martinez, I. (1997). Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. Nucleic Acids Research, 25(22), 4692–4693.

Andújar, C., Arribas, P., Yu, D. W., Vogler, A. P., & Emerson, B. C. (2018). Why the COI barcode should be the community DNA metabarcode for the metazoa. Molecular Ecology, 27(20), 3968–3975.

Baird, D. J., Pascoe, T. J., Zhou, X., & Hajibabaei, M. (2011). Building freshwater macroinvertebrate DNA-barcode libraries from reference collection material: formalin preservation vs specimen age. Journal of the North American Benthological Society, 30(1), 125–130.

Beng, K. C., Tomlinson, K. W., Shen, X. H., Surget-Groba, Y., Hughes, A. C., Corlett, R. T., & Slik, J. W. F. (2016). The utility of DNA metabarcoding for studying the response of arthropod diversity and composition to land-use change in the tropics. Scientific Reports, 6, 24965.

Binladen, J., Gilbert, M. T. P., Bollback, J. P., Panitz, F., Bendixen, C., Nielsen, R., & Willerslev, E. (2007). The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. PloS One, 2(2), e197.

Bisanti, M., Ganassi, S., & Mandrioli, M. (2009). Comparative analysis of various fixative solutions on insect preservation for molecular studies. Entomologia Experimentalis et Applicata, 130, 290–296.

Blackman, R. C., Mächler, E., Altermatt, F., Arnold, A., Beja, P., Boets, P., … Deiner, K. (2019). Advancing the use of molecular methods for routine freshwater macroinvertebrate biomonitoring – the need for calibration experiments. Metabarcoding and Metagenomics, 3, e34735.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). obitools: a unix-inspired software package for DNA metabarcoding. Molecular Ecology Resources, 16(1), 176–182.

Brandon-Mong, G. J., Gan, H. M., Sing, K. W., Lee, P. S., Lim, P. E., & Wilson, J. J. (2015). Dna Metabarcoding of Insects and Allies: an Evaluation of Primers and Pipelines. Bulletin of Entomological Research, 105(06), 717–727.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., ... & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. The R Journal, 9(2), 378–400.

Bucklin, A., Lindeque, P. K., Rodríguez-Ezpeleta, N., Albaina, A., & Lehtiniemi, M. (2016). Metabarcoding of marine zooplankton: prospects, progress and pitfalls. Journal of Plankton Research, fbw023–fbw028.

Carew, M. E., Coleman, R. A., & Hoffmann, A. A. (2018). Can non-destructive DNA extraction of bulk invertebrate samples be used for metabarcoding? PeerJ, 6(1702), e4980.

Carew, M. E., Metzeling, L., St Clair, R., & Hoffmann, A. A. (2017). Detecting invertebrate species in archived collections using next-generation sequencing. Molecular Ecology Resources, 17(5), 915–930.

Clarke, L. J., Soubrier, J., Weyrich, L. S., & Cooper, A. (2014). Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. Molecular Ecology Resources, 14(6), 1160–1170.

Cognato, A. I., & Vogler, A. P. (2001). Exploring Data Interaction and Nucleotide Alignment in a Multiple Gene Analysis of Ips (Coleoptera: Scolytinae). Systematic Biology, 50(6), 758–780.

Collins, R. A., Bakker, J., Wangensteen, O. S., Soto, A. Z., Corrigan, L., Sims, D. W., ... & Mariani, S. (2019). Non-specific amplification compromises environmental DNA metabarcoding with COI. Methods in Ecology and Evolution, 10(11), 1985–2001.

Cowart, D. A., Pinheiro, M., Mouchel, O., Maguer, M., Grall, J., Miné, J., & Arnaud-Haond, S. (2015). Metabarcoding Is Powerful yet Still Blind: A Comparative Analysis of Morphological and Molecular Surveys of Seagrass Communities. PloS One, 10(2), e0117562–26.

Creedy, T. J., Ng, W. S., & Vogler, A. P. (2019). Toward accurate species-level metabarcoding of arthropod communities from the tropical forest canopy. Ecology and Evolution, 9(6), 3105–3116.

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. Biology Letters, 10(9), 20140562–20140562.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., ... & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. Molecular Ecology, 26(21), 5872–5895.

de Mendiburu, F. (2014). Agricolae: statistical procedures for agricultural research. R Package Version, 1(1).

Dopheide, A., Xie, D., Buckley, T. R., Drummond, A. J., & Newcomb, R. D. (2018). Impacts of DNA extraction and PCR on DNA metabarcoding estimates of soil biodiversity. Methods in Ecology and Evolution, 71(1), 8966–8914.

Drummond, A. J., Newcomb, R. D., Buckley, T. R., Xie, D., Dopheide, A., Potter, B. C. M., ... & Nelson, N. (2015). Evaluating a multigene environmental DNA approach for biodiversity assessment. GigaScience, 1–20.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics , 26(19), 2460–2461.

Elbrecht, V., Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Hajibabaei, M., Wright, M., ... & Steinke, D. (2019). Validation of COI metabarcoding primers for terrestrial arthropods. PeerJ, 7(1), e7745.

Elbrecht, V., & Leese, F. (2017). Validation and Development of COI Metabarcoding Primers for Freshwater Macroinvertebrate Bioassessment. Frontiers of Environmental Science & Engineering in China, 5, 314–311.

Elbrecht, V., Taberlet, P., Dejean, T., Valentini, A., Usseglio-Polatera, P., Beisel, J.-N., ... & Leese, F. (2016). Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. PeerJ, 4(4), e1966–12.

Epp, L. S., Boessenkool, S., Bellemain, E. P., Haile, J., Esposito, A., Riaz, T., … Brochmann, C. (2012). New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. Molecular Ecology, 21(8), 1821–1833.

Erdozain, M., Thompson, D. G., Porter, T. M., Kidd, K. A., Kreutzweiser, D. P., Sibley, P. K., … Hajibabaei, M. (2019). Metabarcoding of storage ethanol vs. conventional morphometric identification in relation to the use of stream macroinvertebrates as ecological indicators in forest management. Ecological Indicators, 101, 173–184.

Ficetola, G. F., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessière, J., … Pompanon, F. (2010). An In silico approach for the evaluation of DNA barcodes. BMC Genomics, (11), 434.

Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Molecular Marine Biology and Biotechnology, 3(5), 294–299.

Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. Nature Communications, 1–11.

Geller, J., Meyer, C. P., Parker, M., & Hawk, H. (2013). Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. Molecular Ecology Resources, 13(5), 851–861.

Gibson, J. F., Shokralla, S., Curry, C., Baird, D. J., Monk, W. A., King, I., & Hajibabaei, M. (2015). Large-Scale Biomonitoring of Remote and Threatened Ecosystems via High-Throughput Sequencing. PloS One, 10(10), e0138432–15.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A. C., & Baird, D. J. (2011). Environmental Barcoding: A Next-Generation Sequencing Approach for Biomonitoring Applications Using River Benthos. PloS One, 6(4), e17497–7.

Hajibabaei, M., Smith, M. A., Janzen, D. H., Rodriguez, J. J., Whitfield, J. B., & Hebert, P. D. N. (2006). A minimalist barcode can identify a specimen whose DNA is degraded. Molecular Ecology Notes, 6(4), 959–964.

Hajibabaei, M., Spall, J. L., Shokralla, S., & van Konynenburg, S. (2012). Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. BMC Ecology, 12(1), 1–1.

Hallmann, C. A., Sorg, M., Jongejans, E., Siepel, H., Hofland, N., Schwan, H., … de Kroon, H. (2017). More than 75 percent decline over 27 years in total flying insect biomass in protected areas. PloS One, 12(10), e0185809.

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. Proceedings of the Royal Society B: Biological Sciences, 270(1512), 313–321.

Hebert, P. D. N., Penton, E. H., Burns, J. M., Janzen, D. H., & Hallwachs, W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. Proceedings of the National Academy of Sciences of the United States of America, 101(41), 14812–14817.

Hebert, P. D. N., & Ratnasingham, S. (2007). BOLD: the barcode of life data system. Molecular Ecology Notes, 7, 355–364.

Hebert, P. D. N., Ratnasingham, S., & de Waard, J. R. (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proceedings of the Royal Society B: Biological Sciences, 270(Suppl_1), S96–S99.

Heeger, F., Bourne, E. C., Baschien, C., Yurkov, A., Bunk, B., Spröer, C., ... & Monaghan, M. T. (2018). Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments. Molecular Ecology Resources, 18(6), 1500–1514.

Holman, L. E., de Bruyn, M., Creer, S., Carvalho, G., Robidart, J., & Rius, M. (2019). Detection of introduced and resident marine species using environmental DNA metabarcoding of sediment and water. Scientific Reports, 9(1), 11559.

Horton, D. J., Kershner, M. W., & Blackwood, C. B. (2017). Suitability of PCR primers for characterizing invertebrate communities from soil and leaf litter targeting metazoan 18S ribosomal or cytochrome oxidase I (COI) genes. European Journal of Soil Biology, 80, 43–48.

Hugerth, L. W., Wefer, H. A., Lundin, S., Jakobsson, H. E., Lindberg, M., Rodin, S., ... & Andersson, A. F. (2014). DegePrime, a Program for Degenerate Primer Desing for Broad-Taxonomic-Range PCR in Microbial Ecology Studies. Applied and Environmental Microbiology, 80(16), 5116–5123.

Janzen, D. H., Hallwachs, W., Blandin, P., Burns, J. M., Cadiou, J.-M., Chacon, I., ... & Wilson, J. J. (2009). Integration of DNA barcoding into an ongoing inventory of complex tropical biodiversity. Molecular Ecology Resources, 9(s1), 1–26.

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., … Yu, D. W. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecology Letters, 16(10), 1245–1257.

Ji, Y., Baker, C. C. M., Li, Y., Popescu, V. D., Wang, Z., & Wang, J. (2020). Large-scale Quantification of Vertebrate Biodiversity in Ailaoshan Nature Reserve from Leech iDNA. bioRxiv, 2020.02.10.941336v1.

Ji, Y., Huotari, T., Roslin, T., Schmidt, N. M., Wang, J., Yu, D. W., & Ovaskainen, O. (2020). SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes. Molecular Ecology Resources, 20(1), 256–267.

Karlsson, D., Hartop, E., Forshage, M., Jaschhof, M., & Ronquist, F. (2020). The Swedish Malaise Trap Project: A 15 Year Retrospective on a Countrywide Insect Inventory. Biodiversity Data Journal, 8, e47255.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution, 30(4), 772–780.

Kaunisto, K. M., Roslin, T., Sääksjärvi, I. E., & Vesterinen, E. J. (2017). Pellets of proof: First glimpse of the dietary composition of adult odonates as revealed by metabarcoding of feces. Ecology and Evolution, 7(20), 8588–8598.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., … Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics , 28(12), 1647–1649.

King, J. R., & Porter, S. D. (2004). Recommendations on the use of alcohols for preservation of ant specimens (Hymenoptera, Formicidae). Insectes Sociaux, 51(2), 197–202.

Kraaijeveld, K., De Weger, L. A., Ventayol García, M., Buermans, H., Frank, J., Hiemstra, P. S., & Den Dunnen, J. T. (2015). Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing. Molecular Ecology Resources, 15(1), 8–16.

Krehenwinkel, H., Fong, M., Kennedy, S., Huang, E. G., Noriyuki, S., Cayetano, L., & Gillespie, R. G. (2018). The effect of DNA degradation bias in passive sampling devices on metabarcoding studies of arthropod communities and their associated microbiota. PloS One, 13(1), e0189188–14.

Krehenwinkel, H., Wolf, M., Lim, J. Y., Rominger, A. J., Simison, W. B., & Gillespie, R. G. (2017). Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. Scientific Reports, 1–12.

Deiner, K., Lopez, J., Bourne, S., Holman, L. E., Seymour, M. Grey, E. K., ... & Lodge, D. M. (2018). Optimising the detection of marine taxonomic richness using environmental DNA metabarcoding: the effects of filter material, pore size and extraction method. Metabarcoding and Metagenomics, 2, 1–15.

Lenth, R., Singmann, H., & Love, J. (2018). Emmeans: Estimated marginal means, aka least-squares means. R Package Version, 1(1).

Linard, B., Arribas, P., Andújar, C., Crampton-Platt, A., & Vogler, A. P. (2016). Lessons from genome skimming of arthropod-preserving ethanol. Molecular Ecology Resources, 16(6), 1365–1377.

Lister, B. C., & Garcia, A. (2018). Climate-driven declines in arthropod abundance restructure a rainforest food web. Proceedings of the National Academy of Sciences of the United States of America, 115(44), E10397–E10406.

Liu, M., Clarke, L. J., Baker, S. C., Jordan, G. J., & Burridge, C. P. (2019). A practical guide to DNA metabarcoding for entomological ecologists. Ecological Entomology, 6, e27295v2.

Lynggaard, C., Nielsen, M., Santos Bay, L., Gastauer, M., Oliveira, G., & Bohmann, K. (2019). Vertebrate diversity revealed by metabarcoding of bulk arthropod samples from tropical forests. Environmental DNA, 56(1), 1637–1613.

Macher, J.-N., Vivancos, A., Piggott, J. J., Centeno, F. C., Matthaei, C., & Leese, F. (2018). Comparison of environmental DNA and bulk-sample metabarcoding using highly degenerate cytochrome coxidase I primers. Molecular Ecology Resources, 18(6), 1456–1468.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ, 3, e1420.

Malaise, R. (1937). A new insect-trap. Entomologisk Tidskrift, 58, 148–160.

Mandrioli, M. (2008). Insect collections and DNA analyses: how to manage collections? Museum Management and Curatorship, 23(2), 193–199.

Martin, J. E. H. (1977). The insects and arachnids of Canada. Part 1: Collecting, preparing, and preserving insects, mites, and spiders. Hull: Publication 1643, Research Branch, Canada Department of Agriculture.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 17(1), 10–12.

Martins, F. M. S., Galhardo, M., Filipe, A. F., Teixeira, A., Pinheiro, P., Paupério, J., ... & Beja, P. (2019). Have the cake and eat it: Optimizing nondestructive DNA metabarcoding of macroinvertebrate samples for freshwater biomonitoring. Molecular Ecology Resources, 19(4), 863–876.

Meusnier, I., Singer, G. A. C., Landry, J.-F., Hickey, D. A., Hebert, P. D. N., & Hajibabaei, M. (2008). A universal DNA mini-barcode for biodiversity analysis. BMC Genomics, 9(1), 214–214.

Meyer, C. P., & Paulay, G. (2005). DNA Barcoding: Error Rates Based on Comprehensive Sampling. PLoS Biology, 3(12), e422–10.

Morinière, J., Cancian de Araujo, B., Lam, A. W., Hausmann, A., Balke, M., Schmidt, S., ... & Haszprunar, G. (2016). Species Identification in Malaise Trap Samples by DNA Barcoding Based on NGS Technologies and a Scoring Matrix. PloS One, 11(5), e0155497–14.

Nielsen, M., Gilbert, M. T. P., Pape, T., & Bohmann, K. (2019). A simplified DNA extraction protocol for unsorted bulk arthropod samples that maintains exoskeletal integrity. Environmental DNA, 1(2), 144–145.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R. B. (2013). Package "vegan." Community Ecology Package, Version, 2(9), 1–295.

Piper, A. M., Batovska, J., Cogan, N. O. I., Weiss, J., Cunningham, J. P., Rodoni, B. C., & Blacket, M. J. (2019). Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. GigaScience, 8(8), 10–22.

Pompanon, F., Coissac, E., & Taberlet, P. (2011). Metabarcoding une nouvelle façon d'analyser la biodiversité: Génomique environnementale: faire parler l'invisible. Biofutur, (319), 30–32.

R Development Core Team (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing

Riaz, T., Shehzad, W., Viari, A., Pompanon, F., Taberlet, P., & Coissac, E. (2011). ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. Nucleic Acids Research, 39(21), e145.

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ, 4, e2584.

Shaw, J. L. A., Clarke, L. J., Wedderburn, S. D., Barnes, T. C., Weyrich, L. S., & Cooper, A. (2016). Comparison of environmental DNA metabarcoding and conventional fish survey methods in a river system. Biological Conservation, 197(C), 131–138.

Shokralla, S., Porter, T. M., Gibson, J. F., Dobosz, R., Janzen, D. H., Hallwachs, W., ... & Hajibabaei, M. (2015). Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. Scientific Reports, 5, 9687–9687.

Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H., & Flook, P. (1994). Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. Annals of the Entomological Society of America, 87(6), 651–701.

Statisticat LLC. 2018. LaplacesDemon: Complete Environment for Bayesian Inference. Bayesian- Inference.com. R package version 16.1.1. https://web.archive.org/web/20150206004624/http://ww w.bayesian-inference.com/software.

Stein, E. D., White, B. P., Mazor, R. D., Miller, P. E., & Pilgrim, E. M. (2013). Evaluating Ethanol-based Sample Preservation to Facilitate Use of DNA Barcoding in Routine Freshwater Biomonitoring Programs Using Benthic Macroinvertebrates. PloS One, 8(1), e51273–e51277.

Taberlet, P., Coissac, E., Hajibabaei, M., & Riesenberg, L. H. (2012). Environmental DNA. Molecular Ecology, 21, 1789–1793.

Telfer, A., Young, M., Quinn, J., Perez, K., Sobel, C., Sones, J., ... & deWaard, J. R. (2015). Biodiversity inventories in high gear: DNA barcoding facilitates a rapid biotic survey of a temperate nature reserve. Biodiversity Data Journal, 3, e6313–176.

Thomsen, P. F., & Sigsgaard, E. E. (2019). Environmental DNA metabarcoding of wild flowers reveals diverse communities of terrestrial arthropods. Ecology and Evolution, 9(4), 1665–1679.

Vesterinen, E. J., Ruokolainen, L., Wahlberg, N., Peña, C., Roslin, T., Laine, V. N., ... & Lilley, T. M. (2016). What you need is what you eat? Prey selection by the bat Myotis daubentonii. Molecular Ecology, 25(7), 1581–1594.

Wangensteen, O. S., Palacín, C., Guardiola, M., & Turon, X. (2018). DNA metabarcoding of littoral hard-bottom communities: high diversity and database gaps revealed by two molecular markers. PeerJ, 6(Suppl S9), e4705–e4730.

Wilson, J.-J., Brandon-Mong, G.-J., Gan, H.-M., & Sing, K.-W. (2019). High-throughput terrestrial biodiversity assessments: mitochondrial metabarcoding, metagenomics or metatranscriptomics? Mitochondrial DNA. Part A, DNA Mapping, Sequencing, and Analysis, 30(1), 60–67.

Yang, C., Wang, X., Miller, J. A., de Blécourt, M., Ji, Y., Yang, C., ... & Yu, D. W. (2014). Using metabarcoding to ask if easily collected soil and leaf-litter samples can be used as a general biodiversity indicator. Ecological Indicators, 46, 379–389.

Yeo, D., Srivathsan, A., & Meier, R. (2020). Longer is not always better: Optimizing barcode length for large-scale species discovery and identification. Systematic Biology.

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., & Ding, Z. (2012). Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods in Ecology and Evolution, 3(4), 613–623.

Zizka, V. M. A., Leese, F., Peinert, B., & Geiger, M. F. (2019). DNA metabarcoding from sample fixative as a quick and voucher-preserving biodiversity assessment method. Genome, 62(3), 122–136.

SVENSK SAMMANFATTNING

Insekter är en av de största och mest artrika djurgrupperna, och de inkluderar många värdefulla ekologiska indikatorarter. Men just för att de är så artrika och delvis dåligt kända är många insekter svåra att kartlägga taxonomiskt eller använda i miljöövervakningen. Det är lätt att samla in stora mängder insekter, men sortering och bestämning av materialet kräver mycket tid och resurser, förutom medverkan av många taxonomiskt kunniga experter.. Metabarkodning ("metabarcoding" på engelska) är en metod som kan lösa dessa problem genom att på genetisk väg snabbt och effektivt artbestämma alla insektsarter i ett miljöprov. Metoden bygger på sekvensering av artspecifika genetiska markörer, vilka först amplifieras med polymeraskedjereaktionen (PCR). Därefter analyseras hela provet med de senaste DNA-sekvenseringsteknikerna, vilka genererar miljontals läsningar av olika DNA-strängar i en och samma körning. Metoder för metabarkodning av insektsprover är dock fortfarande dåligt utvecklade. I den här avhandlingen optimerade jag metabarkodningsmetoder för att inventera insekter i terrestra ekosystem.

I **papper I** designade vi nya PCR-primrar för mitokondriella markörer och utvärderade dem mot befintliga primrar med hjälp av datorbaserade metoder. Vi visade att den bästa markören för metabarkodning av insekter är 16S på grund av dess breda taxonomiska täckning och låga amplifieringsbias. Det finns emellertid signifikant mer referensdata för COI, och dess taxonomiska täckning är tillfredsställande om man använder tillräckligt degenererade primrar (blandningar av primersekvenser). I **papper II** applicerade vi metabarkodning med 16S och med COI på tre olika typer av prover från samma insektssamhällen: Malaisefälleprover (antingen prover av etanolen i vilken proverna förvaras, eller homogenat av de insamlade insekterna) och jordprover. Resultaten visar att analyser med två olika markörer ökar detekteringen av arter i provet jämfört med analyser med en enda markör. De visar också att 16S är bättre än COI för metabarkodning av eDNA-prover (prover av fritt förekommande DNA i miljön, till exempel i jordprover eller vattenprover), eftersom de mindre degenererade 16S-primrarna inte amplifierar lika många arter i ovidkommande organismgrupper som svampar och bakterier. Slutligen visar resultaten att analyser av homogenat och av konserveringsvätskan från Malaisefälleprover ger slående olika resultat. Stora och kraftigt sklerotiserade insekter läcker inte DNA till konserveringsvätskan i någon nämnvärd omfattning till skillnad från små och svagt sklerotiserade insekter, men deras DNA tenderar att överskugga de senare insektgruppernas DNA i homogeniserade prover. I **papper III** utvärderade vi prestandan hos olika metoder för icke-destruktiv extraktion och rening av DNA. Vi utsatte prover av olika artificiellt konstruerade insektssamhällen för inkubation i antingen en mild eller en mer aggressiv extraktionsbuffert under kort eller lång tid. DNA extraherades sedan med en manuell eller en automatiserad metod . Vi fann att den milda bufferten och den kortare inkubationstiden bevarade insekternas morfologi bäst, samtidigt som de gav de bästa DNA-metabarkodningsresultaten; eningsmetoden hade liten eller ingen effekt på resultaten. Slutligen, i **papper IV**, testade vi det allmänna antagandet att höga koncentrationer av etanol, även om de är optimala för att bevara DNA, gör insekter ömtåliga och svåra att studera morfologiskt. Vi konserverade insekter i etanol med olika koncentration och utsatte dem för olika påfrestningar, till exempel skakning (vortex) eller olika typer av transport. Våra resultat bekräftar att höga koncentrationer av etanol gör insekter mer bräckliga, även om effekten är mindre uttalad hos robusta insekter som skalbaggar och myror. Våra resultat indikerar också att det är riskfritt att skicka prover med post om insekterna förvaras i måttligt koncentrerad etanol (70 %).

Forskningen som presenteras i den här avhandlingen kan sammanfattningsvis sägas utgöra ett betydande steg framåt i utvecklingen av metoder för att bevara och genetiskt analysera prover av insekter från terrestra ekosystem för taxonomisk och ekologisk forskning och för olika typer av miljöövervakningsprogram.

RESUMEN EN ESPAÑOL

Los insectos son uno de los grupos animales más abundantes y diversos, incluyendo muchas especies de alto valor como indicadores ecológicos, pero los proyectos de descubrimiento de especies y de mapeo de biodiversidad focalizados en este grupo suponen un gran reto. Aunque usando dispositivos de captura masiva se pueden recolectar grandes cantidades de insectos, la labor de identificación de las especies encontradas es un proceso arduo y que requiere muchos recursos (tanto humanos como materiales). Usando metabarcoding, la técnica de secuenciación masiva de marcadores genéticos específicos de cada especie amplificados en una PCR desde muestras ambientales, se podría solucionar este problema. Sin embargo, esta técnica está aún dando sus primeros pasos. En esta tesis he optimizado los métodos de metabarcoding para su aplicación en inventarios de especies y para acelerar el proceso de descubrimiento taxonómico de insectos terrestres.

En el **artículo I** diseñamos nuevos primers para genes mitocondriales y los comparamos con otros ya publicados usando simulaciones por ordenador. Demostramos que el mejor marcador para metabarcoding de insectos es el 16S, gracias a su amplia cobertura taxonómica y mínimo sesgo de amplificación. Sin embargo, la cantidad de material de referencia para este marcador es con mucha diferencia inferior a la que existe para COI, y la cobertura taxonómica de este último es razonablemente similar a 16S cuando se usan primers suficientemente degenerados (mezclas de primers con distintas secuencias). En el **artículo II** utilizamos metabarcoding de 16S y COI con diferentes tipos de muestras de la misma comunidad de insectos: capturas de trampas de Malaise (tanto el alcohol que se usa como preservante como las muestras homogeneizadas) y muestras de suelo. Los resultados evidencian que usar la estrategia de dos marcadores incrementa la detección de biodiversidad sobre la estrategia de un solo marcador. También mostramos que 16S es mejor que COI cuando se utiliza con muestras ambientales, debido a que los primers están menos degenerados y no amplifican el ADN de organismos fuera del grupo de interés tanto como los de COI. Por último, demostramos que el análisis del tejido homogeneizado y del alcohol de las trampas producen resultados muy diferentes. Los insectos de mayor tamaño y con cutículas más esclerotizadas liberan menos ADN al alcohol que los insectos pequeños y de cutículas blandas, pero los primeros contribuyen proporcionalmente mucho más que los segundos al conjunto total del ADN en las muestras homogeneizadas. En el **artículo III** evaluamos el rendimiento de varios tratamientos de lisis leve y no destructiva, así como varios métodos de purificación del ADN. Para ello incubamos muestras de comunidades artificiales en una solución de digestión mod-

erada o en una más agresiva químicamente, durante un periodo de tiempo más corto o más prolongado. El ADN fue posteriormente extraído usando un protocolo manual o uno automatizado en un robot de laboratorio. Descubrimos que una digestión más ligera durante menos tiempo era capaz de preservar mejor la morfología de los insectos y al mismo tiempo producir los resultados de metabarcoding más precisos. El método de purificación del ADN, además, no tiene ningún efecto sobre los resultados. Por último, en el **artículo IV** investigamos la extendida noción de que el alcohol en concentraciones elevadas induce fragilidad en los insectos y los hace difíciles de manejar. Preservamos insectos en distintas concentraciones de alcohol y los sometimos a varios procesos potencialmente dañinos, como sacudidas o transporte. Concluimos que verdaderamente el alcohol en altas concentraciones hace que los insectos se vuelvan frágiles, aunque este efecto es menos pronunciado en insectos de consistencia robusta. Los resultados también indican que enviar las muestras con servicios de mensajería es seguro, siempre que los insectos vayan preservados en concentraciones de alcohol intermedias (70 u 80 %).

En resumen, esta tesis representa un avance significativo en el desarrollo de métodos para la preservación y análisis de muestras de insectos terrestres para estudios de biodiversidad, monitoreo ambiental e investigación taxonómica.

ACKNOWLEDGEMENTS

First of all, I want to thank Fredrik, my supervisor, for giving me the opportunity and the trust to work on this project that had been on his head for a while before I started. Specially, the way of asking for more analyses (even the evening before submitting a paper), when instead of “Can you do ...?” you said “Why don’t you do ... . I’m sure you can work that out.” really gave me the boost and confidence that I could deal with the challenges of this project. I don’t know if I succeeded, but I tried my best.

I want to also thank my co-supervisors, Niclas and Johannes, for always being willing to help me with their expertise when I needed it. Not only for teaching me how to set up a Malaise trap, or helping me with the first DNA extractions from soil, but also for your ideas and discussions. In the same line, I am very grateful to my co-authors: Rodrigo (thanks for all those hours of coffee and blowing off steam), Tomas (thanks for your support as well!), Anders and Piotr. And to those that are co-authors, but it is not down in paper yet.

Thanks of course to the members of the Ronquist lab and the IBA team, Andreia, Ela, Piotr, Scarlett, for their ideas, discussions and help. And, as important as the scientific team, many thanks to the human team: Allison, Mariana, Viktor, Erik, Miroslav, David, Johanna, Johannes, Edana, Marianne, Petter, Nora, Dave, Patricia, Tatiana, Nick, Tom, Mozes, Erik, Alice. A special mention goes to all the BIG4 team as well, for welcoming a flatworm taxonomist in their lines and open my eyes to the wonders of entomology. We had a lot of fun in those workshops, I hope we can all meet again someday.

Finally, I want to thank my family, for believing in me since I was a lonely kid searching for spiders in the school’s walls and memorizing the animal encyclopaedias. Thanks to my sister for all her support and fun when we talk about science, and to her and Sera for being home in a foreign land. Speaking of home, thanks to my wife for moving her home to Sweden with me, and accompanying me in every step of the way.